Monthly Rainfall / Water Discharge of Cervik A
Project II

Jeremy Morris

October 21, 2005

For this project, we will be using a simple linear model for our data. We will define the linear model as

$$Y_i = \beta_i X_i + \alpha_i \tag{1}$$

For each of $i = 5$ months in the data set, the values for Y_i will be the observed water discharges for the i^{th} month. Similarly, the X_i will be the observed rainfalls. What we hope to do is to find estimators for the slope and intercept of each line, $\hat{\beta}$ and $\hat{\alpha}$. This will be done using the usual linear least squares formula discussed in class. All calculations are done using the various functions and methods available in the `R` programming language.

1 Plots for each month

Figure (1) shows a model fit to the data for each month given in the original data. There are a few exceptional outliers. For example, the water discharge for October 1971 affects the model terribly. This outlier will be removed for the rest of the analysis. A comparison plot can be seen in Figure (2). Here, we see a plot of the original data along with a plot of the data with the outlier removed. The line appears to fit better without the outlier. The model for June does not appear to match the data as well as for the other months. The data for 1971 will be removed, this point seems to be the worst outlier. The effect of removing this data point can be seen in Figure (3).

2 How well do the models fit?

Often R^2 is used as a measure of how well the model fits the data. The popular interpretation is that R^2 is the fraction of variance expliained by the model. The R^2 coefficient is calculated using the formula

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} \tag{2}$$

The interpritation of R^2 follows from this formula. The equation is the ratio of the fitted variance with the observed variance in the dependant variable.

We can also use an F test to determine whether the model is significant or not. In this case we want to test

$$H_0 : \beta_1 = 0 \qquad H_a : \beta_1 \neq 0 \tag{3}$$

We will calculate the test statistic

$$F = \frac{(n-2)R^2}{1-R^2} \tag{4}$$

And reject H_0 when $F > F^{\alpha}_{1,n-1}$. From our text, we note that the rejection of H_0 means that there is a significant regression and the x values cannot be ignored. It does not mean that the model is very adequate for prediction purposes. A working rule suggests that the F statistic be at least four or five times $F^{\alpha}_{1,n-1}$ to provide an adequate model for predicion.

Table (1) displays the R^2, F and $F^{\alpha}_{df_1,df_2}$ values along with the values for the degrees of freedom (df_1, df_2) with $\alpha = 0.05$. This table seems to indicate that the monthly models fit

Month	R^2 value	F	$F^{\alpha}_{df_1,df_2}$	df_1	df_2
June	0.569	26.4	4.35	1	20
July	0.808	88.6	4.32	1	21
Aug	0.851	120.4	4.32	1	21
Sept	0.641	37.5	4.32	1	21
Oct	0.736	55.8	4.35	1	20

Table 1: Hypothesis tests for each model by month

the data adequately. We see that the model for June fits the worst, which is to be expected since a look at the plot also indicates a bad fit. We also see that the model for the months of July and August seem to fit the best. This could be because those months are the most stable being in the middle of the summer. The other data occur during seasonal changes, which may explain the presence of more variability.

3 Normally Distributed Errors

Table (2) shows the p-values for the Shapiro-Wilks and Crámer-von Mises tests. Both tests work well in this case, one is parameter invariant and the other is based on the ECDF, therefore we do not have to normalize the data. We see that if we use the null hypothesis that the residuals are normally distributed, we cannot reject for any of the months at the $\alpha = 0.05$ level. Therefore we say that we are 95% sure that the residuals are normally distributed for all months.

Month	Shapiro-Wilks	Cramer-von Mises
June	0.402	0.503
July	0.477	0.131
Aug	0.460	0.512
Sept	0.152	0.139
Oct	0.687	0.576

Table 2: Distribution of Residuals

4 Fitting a model to all the data

4.1 A different model for each month

If there is a different model for each month, we should see a significant difference in the values for the α_i and β_i. The values for α and β are displayed in Table (3), and seem to indicate that the slope for each month may be the same, but the intercepts are different for each month. This would imply that the water discharge is the same relative to the rainfall recieved, but the amount of rainfall is different from month to month. Meaning that we cannot create one model to fit all the data, and have the model remain significantly accurate.

Month	α	β
June	-22.412	0.651
July	-32.576	0.788
Aug	-33.987	0.810
Sept	-17.534	0.682
Oct	-12.693	0.788

Table 3: Coefficients of each model

To make this more precise, we can combine all our data into one model and test the hypothesis that the slopes for each month are the same. We can construct the equation

$$Y = X\gamma \tag{5}$$

where (5) is given by

$$\left(\begin{array}{c} y_{11} \\ y_{12} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ \hline y_{2n_2} \\ \cdots \\ \hline y_{K1} \\ \vdots \\ y_{Kn_K} \end{array}\right) = \left(\begin{array}{cccc} x_{11} & 0 & \cdots & 0 \\ \vdots & 0 & \cdots & 0 \\ x_{1n_1} & 0 & \cdots & 0 \\ 0 & x_{21} & \cdots & 0 \\ 0 & \vdots & \cdots & 0 \\ 0 & x_{2n_2} & \cdots & 0 \\ \hline \cdots & \cdots & \cdots & \cdots \\ \hline 0 & 0 & \cdots & x_{K1} \\ 0 & 0 & \cdots & \vdots \\ 0 & 0 & \cdots & x_{Kn_K} \end{array}\right) \left(\begin{array}{c} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{array}\right) \tag{6}$$

Where there are n_i observations for each month and K months. We need to calculate the F statistic

$$F = \frac{(RSS_{H_0} - RSS_{H_a})/(K-1)}{RSS_{H_a}/(N-2K)} \tag{7}$$

where $N = \sum_{i=1}^{K} n_i$. When we calculate (5) under H_0, we set $\beta_1 = \beta_2 = \cdots = \beta_K = 1$ and use the observed β_i under H_a. Then we reject when $F > F^{\alpha}_{2K,N-2K}$ for $\alpha = 0.05$. By the results in Table (4) we cannot reject H_0. This implies that the slopes of the lines are the same.

F	$F^{\alpha}_{2K,N-2K}$
0.043	1.923

Table 4: F test for parallel lines

As mentioned earlier, we do not necessarily need to test if the intercepts are the same a simple examination of the α values tells us that they will not be. This leads us to the conclusion that one model does not exist for all of the data.

4.2 Modifications to fit a model

To fix the problem of having different models for each month, we could modify the model such that it takes into account the differing amount of rain recieved each month. This means that we use the fitted β value which is the same for each month and scale the model based on a chosen α value.

5 Can the Model Be Improved?

To improve the model for predicting water discharges, all we have to do is think about the situation a little bit. It can easily be argued that the water discharge during the later part of a month will be based on the amount of rain during the month. But, what about the discharge during the beginning of the month? This will mostly be determined by the rainfall and water discharge during the preceding month. In this case our model is altered from (1) to

$$Y_i = \varphi_1 X_i + \varphi_2 X_{i-1} + \varphi_3 Y_{i-} + \varphi_4 \tag{8}$$

If this modification improves our models, then we should see an increase in both the R^2 coefficient and the ratio of F values for each model. Table (5) displays this information.

Month	$R^2_{\{I\}}$	$R^2_{\{O\}}$	$\frac{F_{\{I\}}}{F^\alpha}$	$\frac{F_{\{O\}}}{F^\alpha}$
June	NA	NA	NA	NA
July	0.890	0.808	16.46	20.50
Aug	0.883	0.851	15.33	27.84
Sept	0.743	0.641	5.84	8.67
Oct	0.897	0.736	16.56	12.83

Table 5: Comparisons for imporoving the model

It is interesting that in every case we see an increase in the R^2 coefficient, but not in the F statistic. However, the suggested ratio of higher than four or five is maintained, so that we can still say that these models adequately fit the data.

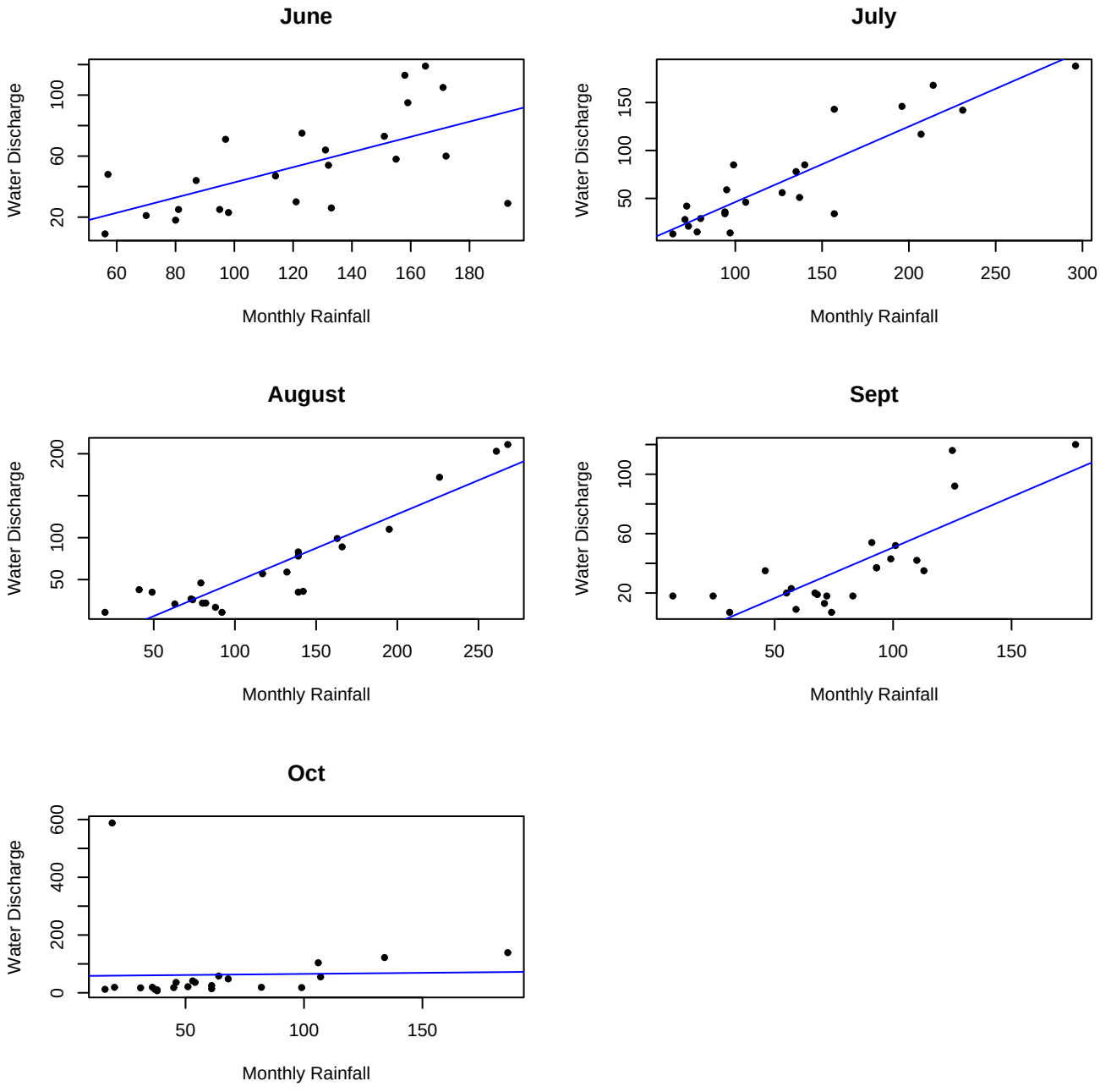

Figure 1: Fitted Lines

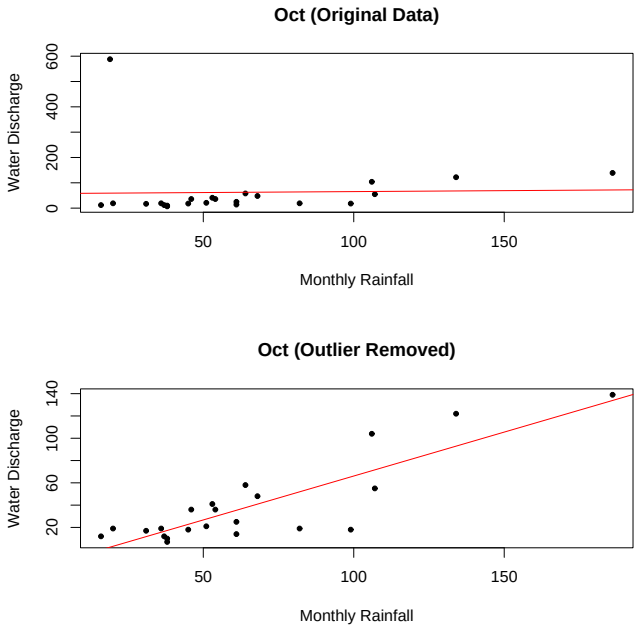

Figure 2: October data with and without outlier

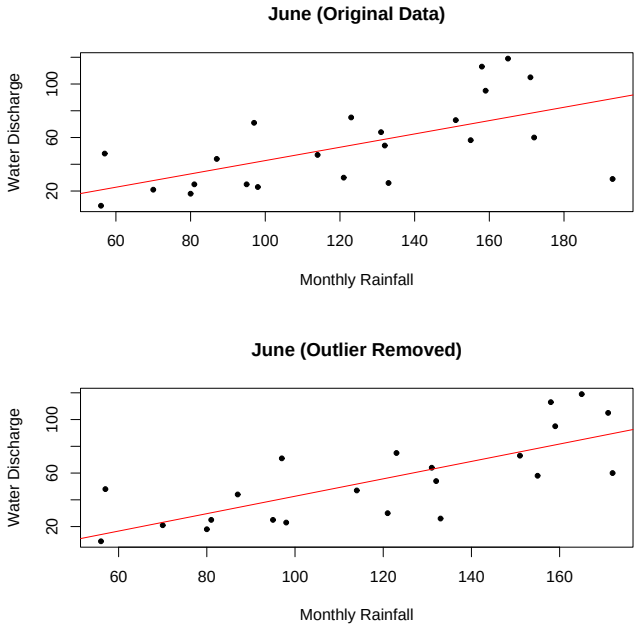

Figure 3: June data with and without outlier